MDPI

Article

T^2 Hotelling control chart for qualitative variables

Wilson Rojas-Preciado[1,2,*], Mauricio J. Rojas-Campuzano[3], Purificación Galindo-Villardón[2], Omar Ruiz-Barzola[3]

[1] Machala Technical University - Machala, Ecuador; wrojas@utmachala.edu.ec
[2] University of Salamanca Salamanca, España; wrojas@usal.es; pgalindo@usal.es
[3] Polytechnic School of the Littoral Guayaquil, Ecuador; maujroja@espol.edu.ec; oruiz@espol.edu.ec
* Correspondence: wrojas@utmachala.edu.ec; Tel.: +593-992-83-3719
† Current address: Updated affiliation

Version March 30, 2023 submitted to Mathematics

Simple Summary: The T2Qv control chart is presented as a multivariate statistical process control technique that performs an analysis of qualitative data through multiple correspondence analysis (MCA), multiple factorial analysis, and the Hotelling T2 chart.

Abstract: The scientific literature is abundant regarding control charts in multivariate environments for numerical and mixed data; however, there are few publications for qualitative data. Qualitative variables provide valuable information on processes in various industrial, productive, and social contexts. Educational processes are no exception and have multiple variables associated with students, teachers, and institutions. When there are many variables, there is a risk of redundant or excessive information, so the application of multivariate methods for dimension reduction to retain a few fictitious variables, a combination of the real ones, that synthesize most of the information is viable. In this context, the T2Qv control chart is presented as a multivariate statistical process control technique that performs an analysis of qualitative data through multiple correspondence analysis (MCA), multiple factorial analysis, and the Hotelling T2 chart. The interpretation of out-of-control points is carried out by comparing MCA charts and analyzing the χ^2 distance between the categories of the concatenated table and those that represent out-of-control points. Sensitivity analysis determined that the T2Qv control chart performs well when working with high dimensions. To test the methodology, an analysis was performed with simulated data and another with real data related to higher education. To facilitate the dissemination and application of the proposal, a reproducible computational package was developed in R, called T2Qv, and is available on The Comprehensive R Archive Network (CRAN).

Keywords: Multivariate; Statistical Process Control; Cualitative; Control Charts; R; T2 Hotelling; Superior Education.

1. Introduction

Statistical control plays a very important role in the continuous improvement of processes, and within it, control charts, which help monitor processes, have been extensively used since their creation by Walter Shewhart [1] until today. From univariate charts, countless proposals have been developed, which incorporated the option of monitoring several variables at once [2,3], thereby opening up the field of multivariate statistical process control (MSPC).

The most well-known options in MSPC are: Hotelling's T2 control chart [4], which could be considered the multivariate version of Shewhart's mean chart; MEWMA [5], which is the multivariate version of the weighted mean chart EWMA [6]; or MCUSUM [7], which is the multivariate version of the cumulative sum control chart CUSUM [8].

Several improvements have been made to these multivariate control charts, such as optimization, analytically determining the optimal values of their parameters [9–11], or heuristically [12]. Another proposal is to work without probabilistic distributions or non-parametric versions [13–15], for continuous or batch processes [2].

All of these multivariate control charts have a quantitative focus, meaning that the monitored variables are essentially quantitative, whether discrete or continuous. Initially, different authors used the Mahalanobis distance [16] for this purpose. Subsequently, for the analysis of a combination of continuous and categorical variables, a chart based on the Gower distance [17] was developed. However, addressing problems such as high correlation between features and in the presence of mixed data required the incorporation of classical multivariate statistical techniques, such as Principal Component Analysis [18], Biplot Methods [19,20], Correspondence Analysis [21], STATIS [22–24], Parallel Coordinates [25], and Cluster Analysis [26].

Among the contributions related to control charts that incorporate multivariate techniques, the STATIS-based chart for monitoring batch processes in nonparametric environments stands out [27]; robust bagplot diagrams using Dual STATIS and Parallel Coordinates [28]; the PCA-based multivariate control chart for mixed data, which applies a combination of Principal Component Analysis and Multiple Correspondence Analysis [29]; the Density-sensitive Novelty Weight control chart (DNW) that uses the k-Nearest Neighbor (kNN) algorithm [14]; the Kernel PCA Mix-based chart [30,31]; the T^2 chart based on a combination of PCA for continuous and qualitative data with outlier detection [32]; and the PCA-based control charts for nonparametric environments [33,34].

However, contributions to the development of multivariate control charts for qualitative variables have not been numerous. In this field, proposals have been developed around the analysis of variables that follow a Poisson distribution and the analysis of multinomial variables. The first proposal was made by Holgate [35], who presented a paper on the bivariate Poisson distribution for correlated variables. This model was used as input in the research of authors such as Chiu and Kuo [36], Lee and Costa [37], Laungrungrong *et al.* [38], and Epprecht *et al.* [39].

Another notable proposal is that of Lu [40], who developed a Shewhart control chart for multivariate processes with qualitative variables, when the quality characteristic takes binary values, which they called a multivariate np (MNP) chart. In the multinomial context, Ranjan-Mukhopadhyay [41] proposed a multivariate control chart using the Mahalanobis D^2 statistic for attributes that follow a multinomial distribution. In addition, for multinomial processes under the fuzzy approach [42], Taleb [43] introduced control charts for monitoring multivariate processes with multidimensional linguistic data, based on two procedures: probability theory and fuzzy theory. Pastuizaca-Fernández *et al.* [44] presented a fuzzy-focused multivariate multinomial T2 control chart.

Saltos Segura *et al.* [45] claim that quality control tools can be considered not only for monitoring industrial processes but also processes related to education, such as student performance evaluation. These authors applied the concept of depth, which transforms a multivariate observation into a univariate index, which is susceptible to monitoring on a control chart, and for this, they used the r chart. They also used cluster analysis to establish thresholds that facilitate the formation of groups and establish student profiles through descriptive measures.

In the study of processes that occur in the social environment, qualitative variables are very frequently used. It's not that quantitative data is absent, but in the databases used for these analyses, nominal and ordinal qualitative variables are abundant, sometimes more so than numeric variables.

López [46] points out that when observing many variables on a sample, it is presumable that some of the collected information may be redundant or excessive. In such cases, multivariate methods for reducing dimensionality attempt to eliminate this information by combining many observed variables to arrive at a few fictitious variables that, while not observed, are a combination of the real variables and synthesize most of the information contained in the data. In this case, the type of variables being handled should be taken into account. If they are quantitative variables, techniques that allow this treatment may be Principal Component Analysis [47,48] or Factor Analysis [49–51], while for

qualitative variables, it is recommended to apply Multiple Correspondence Analysis, Homogeneity Analysis, or Multidimensional Scaling Analysis.

In statistical process control, contributions to the development of control charts for qualitative variables are still in their infancy, with few publications focusing on the analysis of quality characteristics in industrial processes, but not social processes. Upon analyzing the procedures published by the authors cited in this study, limitations are detected that could restrict their application, such as the analysis of few quality characteristics, the use of samples composed of individual elements instead of groups, and the difficulty of working with many categories simultaneously. Thus, the need arises for a control chart for the representation of p qualitative variables that can work with multiple nominal and ordinal categories and facilitate the identification of the causes that can lead to the process being out of control and that can be applied in social processes.

This article addresses the aforementioned limitations regarding control charts for qualitative variables and their application in social environments. For this reason, its objective is to develop a control chart for qualitative variables using multivariate statistical methodologies, to contribute to the diversification of techniques in the phase I of statistical process control.

This article is organized as follows: the Introduction, which establishes the conceptual and referential background of multivariate control charts applied to qualitative variables; section 2, methods, which details the procedure followed in the development of the proposed control chart; section 3 describes the computational complement that facilitates the application of this methodology; section 4 shows the results through the analysis of simulated data; section 5 corresponds to the sensitivity analysis that relates the number of dimensions analyzed versus the reliability of the results. Section 6 presents the discussion through a comparative analysis between the T2Qv control chart and the proposals of other authors. Finally, section 7 establishes the conclusions.

2. Methodology

2.1. Notation

The table 1 contains elements, representation, and examples of how the algebraic elements addressed in the methodology are presented.

Elements	Representation	Example
Scalars	Lowercase letters	v, λ
Vectors	Lowercase bold letters	$\mathbf{v}, \mathbf{u}$
Matrices	Uppercase bold letters	$\mathbf{V}, \mathbf{X}$
Three-way matrices (Data cubes)	Uppercase letters with double stroke	$\mathbb{C}, \mathbb{X}$

Table 1. Algebraic elements

Throughout the article, letters will be used to refer to necessary parameters, which are listed in table 2:

Letter	Meaning	Specification
p	Number of dimensions	
K	Total number of tables (Specifies the depth of the data cube)	
k	Table index	k=1,2,...,K
T	Transpose matrix index	$\mathbf{X}^{\mathbf{T}}$
n	Sample size of the k tables	

Table 2. Notation

2.2. *Multiple Correspondence Analysis (MCA)*

Given that we are working with qualitative variables, Multiple Correspondence Analysis [21] is applied to analyze the similarity between categories [46] based on the χ^2 distance, which is a similar analysis to Principal Component Analysis.

In this case, we are not using the French approach [52], but the Anglo-Saxon approach, where MCA is called Homogeneity Analysis or Dual Scaling. We use the Burt table [21] and start from a data matrix with p qualitative variables, each with h categories $(h > 1)$.

The matrix is composed of the n rows or observations and p columns or variables, where each cell contains one of the aforementioned categories. It is equivalent to the disjunctive matrix $\mathbf{Z}$, which breaks down the variables into each of their modalities and records the occurrence of events in a binary form [21].

The Burt table is given by:

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z} \tag{1}$$

The matrix B in 1 is formed by the absolute frequencies, which are transformed into relative frequencies by dividing the values in the matrix by the total frequency, resulting in the matrix $\mathbf{P}$.

The row and column mass vectors, *mf* and *mc*, respectively, are obtained through the row and column margins of the matrix $\mathbf{P}$.

The standardized residuals matrix $\mathbf{S}$ is then obtained.

$$\mathbf{S} = \mathbf{D_{row}}^{-\frac{1}{2}}(\mathbf{P} - \mathbf{mf}\ \ \mathbf{mc}')\mathbf{D_{column}}^{-\frac{1}{2}} \tag{2}$$

Where $\mathbf{D_{row}}$ is a diagonal matrix containing the row masses and $\mathbf{D_{column}}$ is a diagonal matrix containing the column masses.

Singular value decomposition (SVD) is applied to the matrix $\mathbf{S}$ (Equation 2):

$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{V}' \tag{3}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{D}$ is a diagonal matrix containing the singular values.

Then, the standardized coordinates are obtained by applying Equations 4 and 5.

$$\mathbf{X} = \mathbf{D_{row}}^{-\frac{1}{2}}\mathbf{U} \tag{4}$$

$$\mathbf{Y} = \mathbf{D_{column}}^{-\frac{1}{2}}\mathbf{V} \tag{5}$$

2.2.1. Generalization to k tables

If there are K tables, with the same structure and composed of qualitative variables, what is described in section 2.2 is applied to each of the K tables, obtaining the set of K tables with the initial format.

Each of the K sets of coordinates obtained in the previous step is denoted as C. In order to detect the magnitude of the latent variables, the absolute value of the elements of the matrix $C_k(k = 1, \ldots, K)$ is taken. Thus, a set of K tables of coordinates (loadings) is obtained, whose rows correspond to the observed variables and the columns to the latent variables.

2.2.2. Normalization of tables

Normalization [53] from Multiple Factor Analysis (MFA) is applied to the K tables $\mathbf{C}$.

Let λ_1^k be the first eigenvalue obtained from the singular value decomposition of the k-th table $\mathbf{C}$. The table is normalized by multiplying it by $1/\lambda_1^k$. This results in the table $C^{'}$, which corresponds to

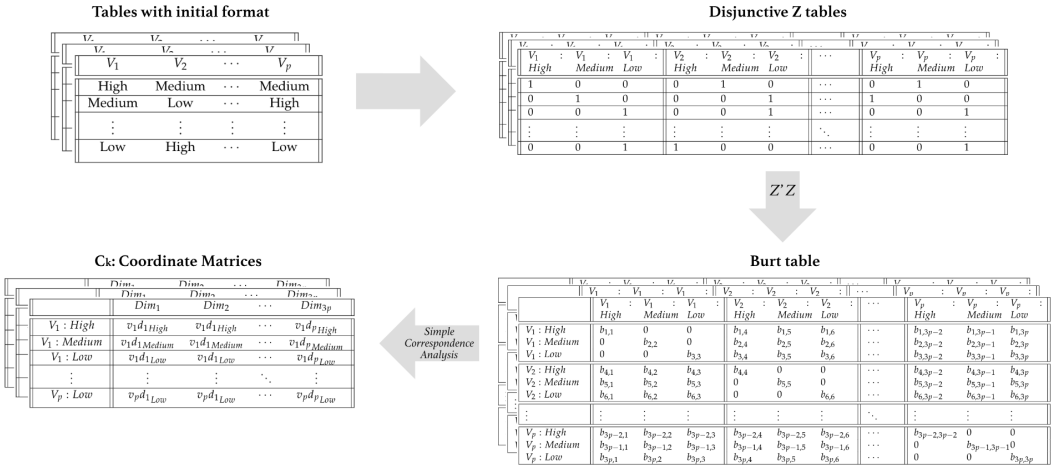

Figure 1. Procedure of MCA for K tables

the normalized coordinate table. Individually, for the case of the k matrix, the following expression would be obtained.

$$\mathbf{C}'_{\mathbf{k}} = \frac{1}{\lambda_k^1}\mathbf{C}_{\mathbf{k}} \tag{6}$$

Up to this point, we have a set of normalized coordinate matrices, whose rows contain the observed variables and columns contain the latent variables.

The expression in equation 6 applied to K tables is represented in Figure 2, which shows the scheme for preparing the tables prior to obtaining centrality vectors used by the multivariate control chart.

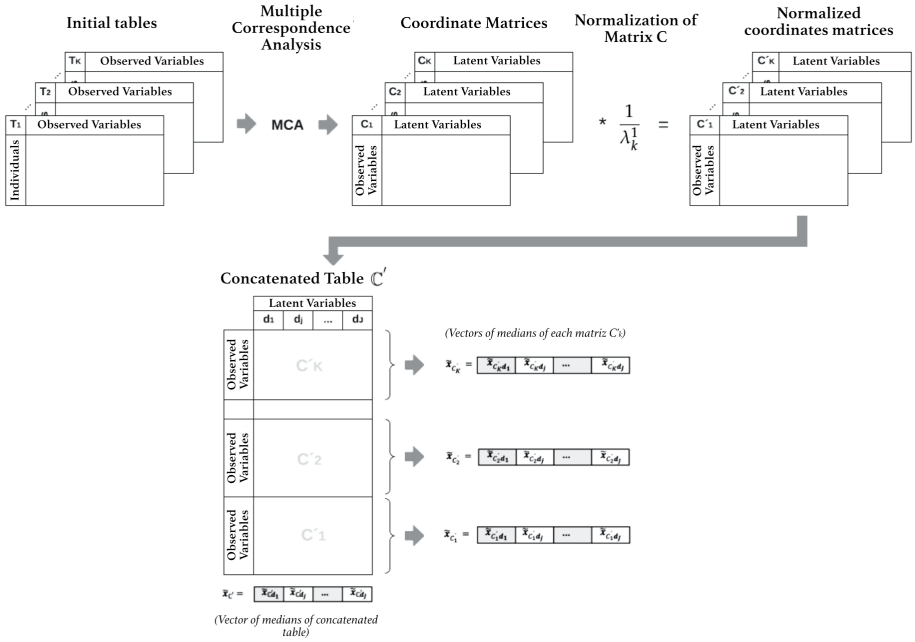

Figure 2. Scheme of the process for obtaining median vectors

By unifying the K normalized tables $C^{'}$ into a single one, we obtain the concatenated matrix $\mathbb{C}^{'}$, which contains all the elements of the K normalized tables.

$$\mathbb{C}^{'} = [\mathbf{C_1^{'}}|\mathbf{C_2^{'}}|,...,|\mathbf{C_K^{'}}]^T \tag{7}$$

The normalization performed by MFA is responsible for weighting the K tables, with the aim of avoiding any imbalance when carrying out the joint analysis of the tables.

From the matrices $\mathbb{C}^{'}$ and $\mathbf{C_k^{'}}$, the median vectors are obtained, as shown in Figure 2.

The vector $\tilde{\mathbf{x}}_{\mathbf{C_k^{'}}}$ explains the central behavior of table k, and the vector $\tilde{\mathbf{x}}_{\mathbb{C}^{'}}$ explains the behavior of the concatenated matrix.

2.3. T2Qv Control Chart

2.3.1. Obtaining the control chart

To define the Hotelling T^2 control chart, the following considerations must be taken into account:

- The matrix $\mathbb{C}^{'}$ (Equation 7) is called Concatenated and serves as a reference for the in-control scenario in phase I of the process control.
- The Hotelling T^2 statistic is usually calculated with the mean vectors and covariance matrix of the in-control process. The proposal of this research is to adopt robustness concepts, using the median vector instead of the mean vector, because medians are not affected by outliers.
- From the concatenated matrix $\mathbb{C}^{'}$, we obtain $\tilde{x_0}$ (median vector of the concatenated matrix) and S_0 (covariance matrix of the concatenated matrix).
- Each matrix $\mathbf{C_k^{'}}$ has the same number of columns.
- The mean vector $\tilde{x_k}$ is tied to the table $\mathbf{C_k^{'}}$, meaning that the control chart will depend on the differences between the matrices $\mathbf{C_k^{'}}$ and the concatenated matrix $\mathbb{C}^{'}$.
- The matrices $\mathbf{C'k}$ follow a multivariate normal distribution with median vector $\tilde{x_k}$ and covariance matrix $\mathbf{S_k}$.

The statistic T^2 is given by:

$$T^2 = n(\mu_k - \mu_0)'\mathbf{\Sigma_0^{-1}}(\mu_k - \mu_0) \tag{8}$$

Taking into account the aforementioned considerations, the statistic T^2_{med} is obtained.

$$T^2_{med} = n(\tilde{x_k} - \tilde{x_0})'\mathbf{\Sigma_0^{-1}}(\tilde{x_k} - \tilde{x_0}) \tag{9}$$

It is known that, under control, T^2 is distributed as a Chi-squared with p degrees of freedom, χ^2_p. In this case, this principle can be applied, as the Concatenated matrix ($\mathbb{C}^{'}$) is used, which represents the in-control scenario.

Since this control chart is based on weighted Mahalanobis distances, it only has an upper control limit. This is given by Equation 10.

$$UCL = \chi^2_{\alpha,p} \tag{10}$$

where p is the number of dimensions and α is the predetermined significance level, with $\alpha = 0.0027$ being considered.

2.3.2. Interpretation of out-of-control points

The multivariate chart for qualitative variables, T2Qv, is capable of indicating that the process has gone out of control, but it does not allow recognizing the causes for this to happen. Each point represented on the chart represents a table (sample), consisting of a group of individuals (observations) and p variables that can have many categories, some of which may exhibit anomalous behavior.

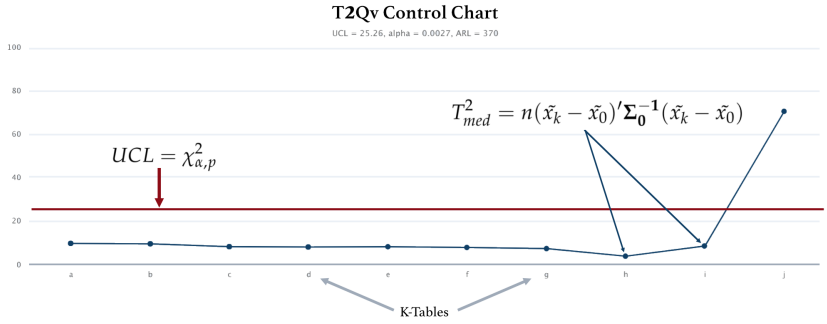

Figure 3. T2Qv Control Chart

Therefore, it is necessary to carefully analyze what is happening with the data of the reported tables to identify the variable(s) that caused the process to go out of control.

This analysis is performed by comparing the location of the points representing the categories of the variables in the MCA of the concatenated table and the location of the points in the MCA charts of each table reported as out of control. The categories that are influencing the out-of-control state are those that show noticeable differences in their location when comparing both tables. To quantify the magnitude of these differences or the anomalous behavior of these categories, the Chi-squared distances between the masses of the columns of the table reported as out of control and the columns of the concatenated table, taken as a reference, are calculated. The higher the value of the statistic, the greater its incidence in the displacement of the centrality of the process that can ultimately lead to an out-of-control state.

3. Computational complement

To facilitate the dissemination and application of the proposed method, a reproducible package has been developed in R. The **T2Qv** package [54] performs the analysis of control of K tables through multivariate control charts for qualitative variables, using the theoretical foundations of multiple correspondence analysis and multiple factor analysis, as well as the conceptual idea of STATIS.

The charts can be displayed in flat or interactive form, and all outputs can be shown in an interactive Shiny panel, and their graphical and numerical results can be exported.

3.1. Description of the T2Qv package

The statistical package T2Qv performs Multiple Correspondence Analysis on the original tables ($\mathbf{T_k}$), generating latent variable matrices ($\mathbf{C_k}$) whose coordinates are subjected to a normalization process, multiplying them by $1/\lambda_1^k$. The normalized coordinate matrices ($\mathbf{T'_k}$) are ordered one below the other, to form a concatenated table ($\mathbb{C}^{'}$), from which the median vector $\tilde{\mathbf{x}}_{\mathbb{C}'}$ is extracted, as well as the median vectors of each matrix $\tilde{\mathbf{x}}_{\mathbf{C'k}}$) that conform it.

With these vectors, the statistics $T^2_{med} = n(\tilde{x_k} - \tilde{x_0})'\mathbf{\Sigma_0^{-1}}(\tilde{x_k} - \tilde{x_0})$ are obtained for each of the analyzed tables, which are represented as points in the T2Qv control chart. Points that fall outside the limit ($UCL = \chi^2_{\alpha,p}$) are reported as out of control.

The T2Qv statistical package allows the interpretation of the anomalous behavior of points outside of control through the comparison of the MCA charts of a table $\mathbf{TC_k}$, which results from concatenating the initial matrices, and each initial table $\mathbf{T'_k}$. The package allows for the selection of the $\mathbf{T'_k}$ tables, so that the researcher can focus their analysis on those identified as out of control.

In addition, the T2Qv package generates an interactive bar chart that represents the χ^2 distances between the column masses of the variables in the $\mathbf{TC_k}$ table and the $\mathbf{T'_k}$ table. Bars denoting greater height identify the variables that are most strongly causing the out-of-control output of the k-th table. This interactive chart includes, through a nested circular chart, a representation of the distribution

of the observed variable categories corresponding to the k-th table, as well as a circular chart of the distribution of categories in the concatenated table ($\mathbf{TC_k}$), facilitating the identification of changes in category distribution.

Thus, the T2Qv package consolidates the methodology proposed in this research and allows for an explanation of when and why the process went out of control.

The functions included in the package and their description are listed in Table 3.

Function	Description
T2 qualitative	Multivariate control chart T2 Hotelling applicable for qualitative variables.
MCAconcatenated	Multiple correspondence analysis applied to a concatenated table.
MCApoint	Multiple correspondence analysis applied to a specific table.
ChiSq variable	Contains Chi square distance between the column masses of the table specified in PointTable and the concatenated table. It allows to identify which mode is responsible for the anomaly in the table in which it is located.
Full Panel	A shiny panel complete with the multivariate control chart for qualitative variables, the two MCA charts and the modality distance table. Within the dashboard, arguments such as type I error and dimensionality can be modified.

Table 3. Functions of the T2Qv package

3.2. Availability

The package is available on the official R repository, The Comprehensive R Archive Network (CRAN), and can be downloaded as follows:

```
install.packages("T2Qv")
```

4. Results

To test the proposed methodology in the Hotelling's T^2 control chart for qualitative variables, an analysis was conducted using simulated data and another using real data applied in the context of higher education. The results were obtained through the application of the T2Qv package.

4.1. Results with simulated data

4.1.1. Simulated data generation

For this study, a simulated database was generated, called *Datak10Contaminated*. It consists of 10 tables, each one composed of 100 rows (observations) and 11 columns, of which the first 10 correspond to the analyzed variables (V1, V2, ..., V10), which contain 3 categories (High, Medium, and Low), while column 11, called *GroupLetter*, contains the classification factor of the groups. For their identification, the tables have been named with the letters of the alphabet, from *a* to *j*. Table *j* has a different distribution than the other nine.

The first 9 tables have their 10 variables with the following distribution:

$$u \sim U[0,1]$$

$$t_{1,..,9} = \begin{cases} Low & si \quad u \leq 1/3 \\ Medium & si \quad 1/3 < u < 2/3 \\ High & si \quad u \geq 2/3 \end{cases}$$

Table *j* or Table 10, in all its 10 variables, follows the distribution presented below:

$$u \sim U[0,1]$$

$$t_j = \begin{cases} Low & si \quad u \leq 1/5 \\ Medium & si \quad 1/5 < u < 2/6 \\ High & si \quad u \geq 2/6 \end{cases}$$

The database is presented in the format established in table 4.

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	GroupLetter
Low	Medium	Medium	High	High	High	Low	Medium	Medium	Medium	a
Low	Low	High	Low	Medium	High	High	High	Low	High	a
High	Medium	High	Low	High	Medium	Medium	High	Medium	Low	a
Medium	Medium	Low	High	Low	Medium	High	Low	Low	High	a
Low	Low	Low	High	Low	High	High	High	Medium	Medium	a
High	High	Medium	Low	High	Low	Medium	Medium	High	Low	a
High	High	Low	Low	Low	Medium	High	Medium	Medium	High	a
Medium	Medium	High	Medium	Medium	High	Medium	High	High	High	a
Low	Low	Low	Medium	High	Medium	Low	Medium	Low	Low	a
Medium	Medium	Medium	High	Low	Medium	High	Low	High	Medium	a

Table 4. Section of the *Datak*10*Contaminated* database.

To verify the difference between the distributions of table 10 and the others, the average of the relative frequencies in the three categories was calculated from table *a* to *i*, for the 10 variables (Annex 1). Then, the average of the mean relative frequencies of the 10 variables was calculated. The result allows to compare the distribution of the categories of the *Datak10Contaminated* table with the theoretical uniform distribution, as shown in table 5.

Categories	Uniform theoretical	Mean of the distributions of the variables in the tables *a*, *b*, ..., *i*	Mean of the distribution of the variables in the table *j*
High	0.333	0.340	0.724
Medium	0.333	0.336	0.092
Low	0.333	0.324	0.184

Table 5. Comparison of the distribution of the categories in the *Datak10Contaminated* table with the theoretical uniform distribution.

The corresponding goodness-of-fit chi-square tests were applied to confirm the distribution of the generated data, as well as the comparison of table *j* with the other tables, confirming significant differences between the distributions (p-value < 0.05), as shown in Appendix 2.

4.1.2. Application of T2Qv package with simulated data

The first result is the graph of Multiple Correspondence Analysis (MCA) applied to the concatenated table (Figure 4). This table is considered the visual reference for the scenario under control for the subsequent analysis of tables reported as out-of-control points in the T2Qv plot.

The MCA reports a total inertia of 63.35%, dimension 1 representing 53.64% of the information, while dimension 2 represents 9.71%. The points in the graph represent the observations of each of the 10 variables at their three levels: *High*, *Medium*, and *Low.* In the MCA plot, observations that are located in the center of the graph represent categories that occur most frequently, while those furthest from the center are rare cases. In this sense, in the concatenated table, there are no observations located at the center of the graph, but they are distributed in groups surrounding the center, which is

explained by the uniform distribution of variable categories in most of the tables, with no one category predominating.

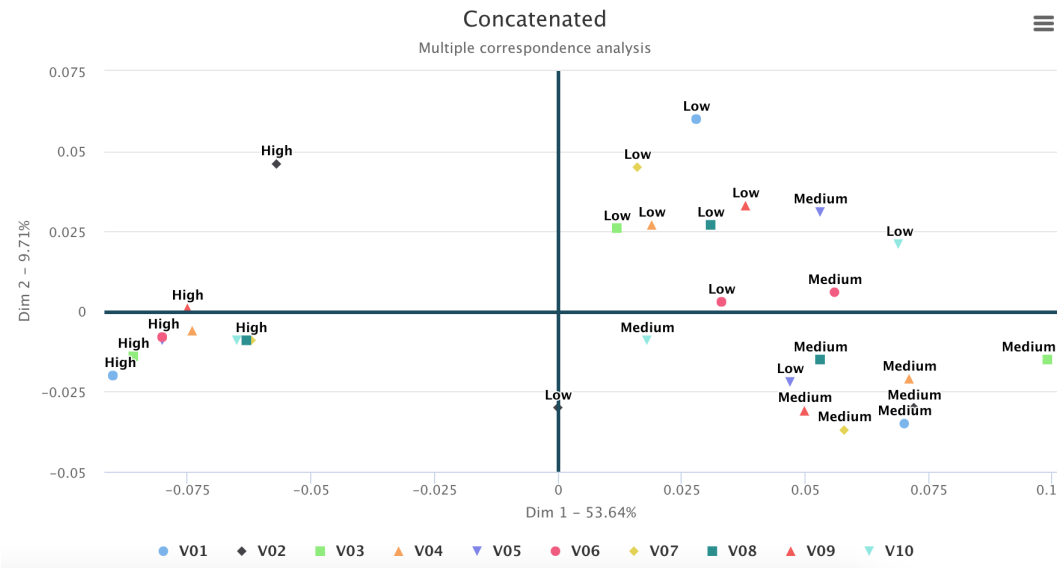

Figure 4. Multiple Correspondence Analysis applied to the concatenated table.

Another result is the MCA applied to a specific table. At this point, one of the arguments that must be taken into account is the selection of the table to be analyzed.

When comparing the plots, it can be observed that the table from point b (5) shows differences with the table from the controlled state (4); however, the differences are not significant enough to generate an out-of-control signal (6, point b).

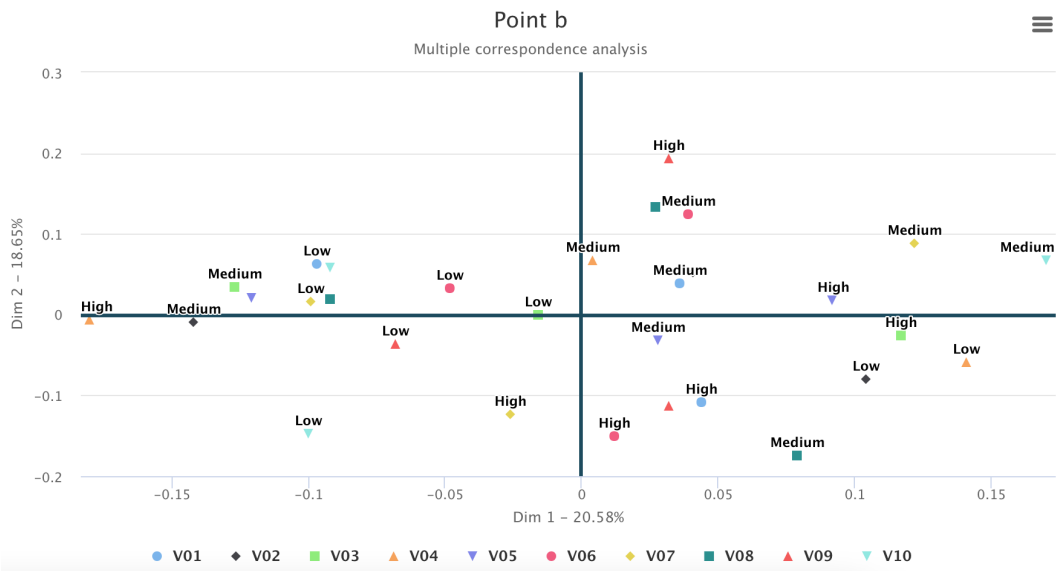

Figure 5. Multiple Correspondence Analysis applied to Table b.

The figure 5 represents the MCA plot of table b, corresponding to a specific moment in the monitored process. This plot, in its two dimensions, represents 39.23% of the information. It is noticeable that the observations in their levels *high*, *medium*, and *low* are randomly distributed in all quadrants of the plot, and a specific pattern of grouping cannot be identified.

The same can be said of the points represented in any of the other tables because they share the same distribution, except for table j, which was designed with a different distribution. However, the use of MCA from figures 5 and 4 still does not allow for detecting whether the process is in control or not. The identification of out-of-control points can be made through the graphical representation of the T^2_{med} statistic, as shown in figure 6.

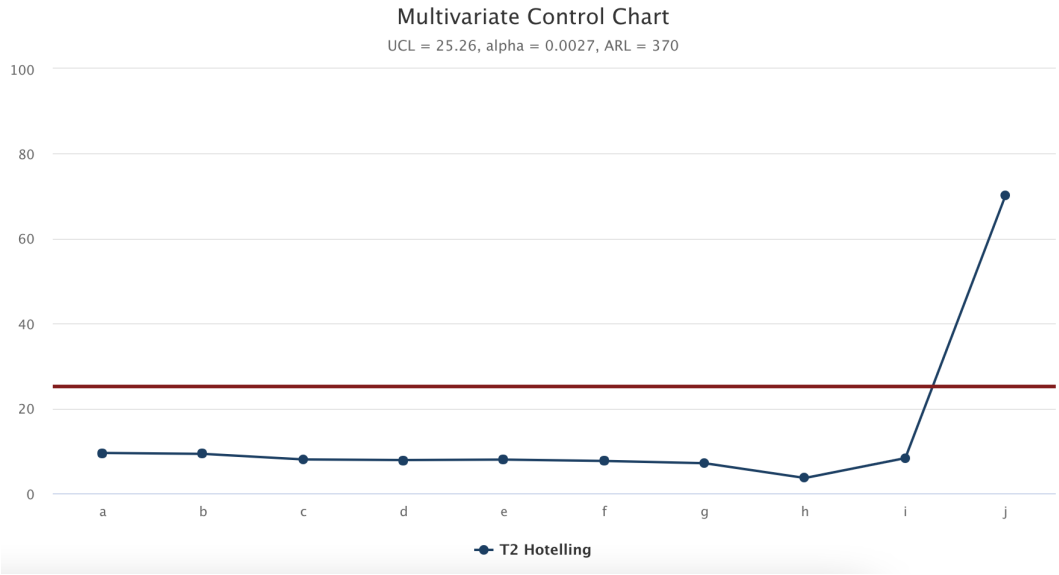

Figure 6. Multivariate control chart T2 Hotelling applied to qualitative variables, *Datak*10*Contaminated*, figure.

The figure 6 presents the T2Qv control chart, based on the adjusted Hotelling's T2 statistic (T^2_{med}), applied to detect anomalies in any of the K analyzed tables. Each of the tables is represented by the points in the chart. A horizontal line representing the upper control limit (UCL) is observed. The lower control limit (LCL) is set to zero.

It is observed that the point representing table j is located above the upper control limit, meaning it has been identified as an out-of-control value. Therefore, it is necessary to analyze the reported data of the table in detail and compare it with the concatenated table to identify the causes of variation and take appropriate actions. To analyze the out-of-control point, a MCA plot of table j is generated and compared to a similar plot of the concatenated table, as presented in figure 7.

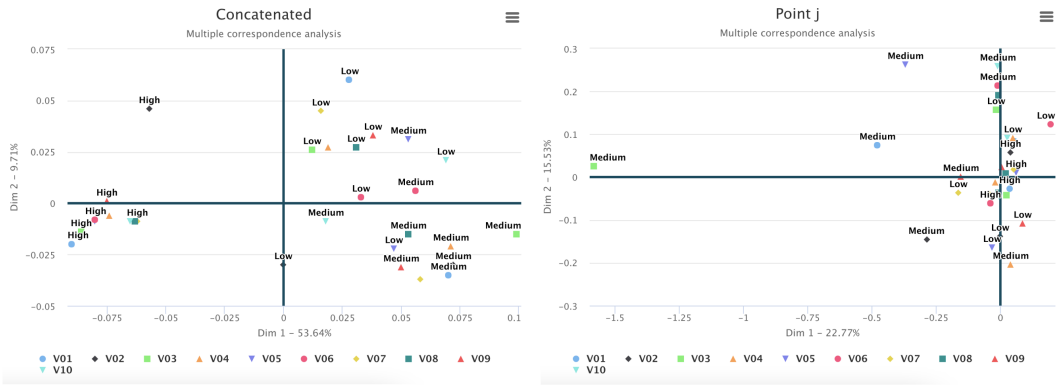

Figure 7. Multivariate control chart T2 Hotelling applicable to qualitative variables, *Datak*10*Contaminated*.

The figure 7 presents the distribution of observations from the concatenated and j tables using MCA plots. The MCA plot of the concatenated table, which serves as the control reference, was already analyzed in Figure 4. The MCA plot of the j table shows a trend of the mean values of the variables to be located on the left side, away from the center of the plot, indicating that the mean values are infrequent. Special attention is deserved by variable 3, which records an observation for the medium level with the farthest value from the group. On the contrary, the high categories have been located at the center, meaning that they are very frequent.

By comparing the plots, it is evident that the data distribution in the MCA plot of the j table is different from the distributions of the other tables and especially from the data distribution in the MCA plot of the concatenated table, which explains why the j point was identified as out of control in the T2Qv plot. This difference is explained in Table 6, which shows the chi-square distance between the observations of the concatenated table and the j table.

Variables	ChiSq
V1	0.06968
V2	0.05010
V3	0.07601
V4	0.04982
V5	0.05205
V6	0.05603
V7	0.03713
V8	0.03702
V9	0.04395
V10	0.06179

Table 6. Chi-square distance between the column masses of table k and the concatenated one, *Datak*10*Contaminated*.

Another way to visualize this information is through a bar chart generated by the T2Qv application (Figure 8).

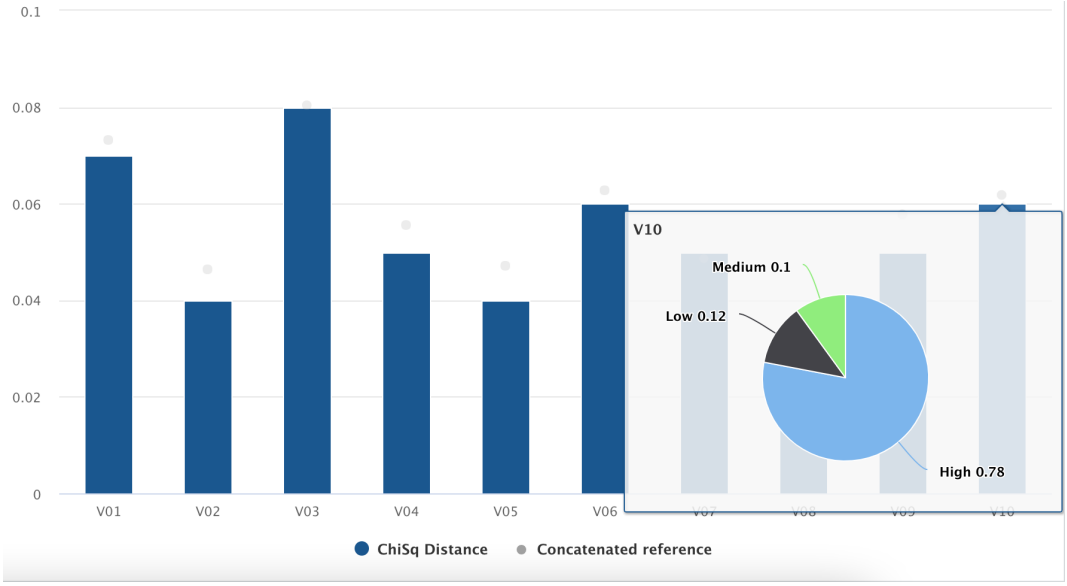

Figure 8. Chi-squared distance between the masses of the concatenated table and the k tables, *Datak*10*Contaminated*.

The bar chart in figure 8 also shows the χ^2 distance between the masses of the concatenated table and those of the k tables in the Datak10Contaminated database, in this case table j. Table 6 shows that

variables V03, V01, and V06 exhibit the highest χ^2 distances between the masses of the concatenated table and table *j* (0.07700, 0.06968, 0.05938), which are represented by the tallest bars in figure 8.

The interactivity of this chart facilitates the observation of the distribution of the variable categories in the analyzed table, and their comparison with the distribution of the variable categories in the concatenated table, as shown in figure 4.1.2.

Figure 9. Distribution of the categories of variables V03, V01, and V06 in the concatenated table and table *j* in the T2Qv application.

Figure presents, in pie charts, the distribution of categories for variables V03, V01, and V06, which recorded the highest Chi-squared distances between the masses of the concatenated table and the *j* table. The charts corresponding to the concatenated table show sectors with equivalent areas, which

is explained by the uniform distribution of the variables, while those of the j table show areas with varying sizes, where the High category has a relatively high frequency in all three cases, and Low has a low frequency. Comparing these charts makes it evident that the distribution of categories presents significant differences between the concatenated table and the j table.

5. Sensitivity Analysis

As mentioned earlier, in the T2Qv plot, an out-of-control point is interpreted as a table (k_i) that includes a quantity or proportion of contaminated variables. In these cases, it is expected that the points on the T2Qv plot will generalize the behavior of these differences in their distribution, thus exceeding the upper control limit (UCL). The location of this control limit varies depending on the number of dimensions represented; when it is high, optimal performance is achieved, while decreasing the number of dimensions that can be represented introduces instability and reduces the reliability of the results.

The proposed control chart is capable of detecting an out-of-control point even with a low number of contaminated variables when working with a high number of dimensions. It is recommended to use $p-1$, where p is the total number of dimensions in the initial matrix (Show in 1). When the number of dimensions is decreased, the height of the upper control limit (UCL) also decreases, resulting in an increased number of out-of-control points, although the variables may not necessarily express significant differences in their values, increasing the probability of obtaining a type I error.

Therefore, the question arises as to how many dimensions can be reduced in the analysis without losing reliability in the result. The importance of this question lies in the need for a reliable chart that identifies out-of-control points even if a dimensionality reduction technique has been applied to the data, without falling into cases of false positives.

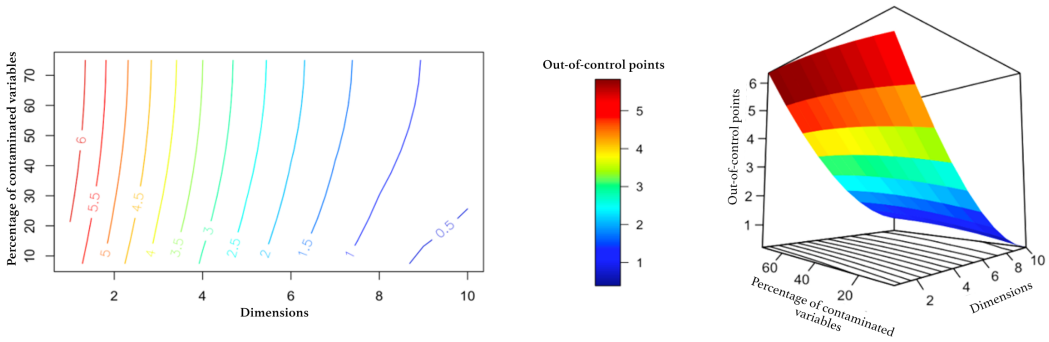

Figure 10. Contour plots and response surface obtained with the T2Qv chart.

The sensitivity analysis uses contour plots and response surfaces (figure 10) to represent the number of out-of-control points, considering the percentage of contaminated variables in the k_i table and the number of dimensions represented. The test data used in the model are recorded in 10 tables, each of which includes 10 variables and each variable has three categories: High, Medium, and Low. Table 10 (or table j) has a different distribution from the others, this being the contaminated table.

It is observed that the model is capable of identifying an out-of-control point when working with p - 1 dimensions (9), even with a low percentage of contaminated variables. When the number of dimensions decreases to p - 2 (8) and the percentage of contaminated variables is close to 100%, it correctly detects 1 out-of-control point. It is also observed that when the number of dimensions is lower, stability is lost and the power of the test is reduced. Consequently, the sensitivity analysis confirms that the T2Qv control chart performs well when working with high dimensions.

6. Discusión

In statistical process control, there are still not many published proposals for control charts for qualitative variables. Differences between procedures for determining statistics and control charts in this field make comparison difficult.

The T2Qv control chart, presented in this article, applies MCA, a multivariate analysis technique that identifies latent structures underlying the set of qualitative data and involves dimension reduction. Therefore, from the outset, a data table with p dichotomous or polytomous variables ($p > 3$) is required. It is worth remembering that the sensitivity analysis determined that this proposal performs well when working with high dimensions and loses stability at low dimensions. Thus, while in the example with simulated data *Datak10Contaminated* presented in this research, the T2Qv analyzes the behavior of 10 variables, in several reviewed studies, the application cases analyze only two or three, which would lead to the application of a simple correspondence analysis, not multiple. Therefore, these cases could not be treated with the T2Qv.

Examples include the Optimal Linear Combination of Poisson Variables for Multivariate SPC, by Epprecht *et al.* [39], whose registered application case in their publication analyzes two variables related to the count of defects in the production of ceramic jars. The GMDS chart by Ali and Aslam [55] was exemplified with a telecommunications dataset taken from Jiang *et al.* [56], consisting of only two variables. The multivariate control chart developed by Pastuizaca-Fernández *et al.* [44]) for p correlated quality attributes applies fuzzy theory and analyzes two tables taken from publications by Taleb [43] and Taleb *et al.* [42], the first with three variables related to frozen food, and the second with three variables on porcelain production.

Another feature of the T2Qv chart is that each sample is a group composed of a set of individuals, a table. The example of simulated data *Datak10Contaminated* includes a set of 10 tables and 11 variables, each table is a sample, consisting of 100 observations and is represented as a point in the Hotelling's T2 chart.

In publications by several authors, it can be observed that in their application examples, only one table is analyzed, with dimensions n (rows) x p (variables), where each row n_i is a sample. For instance, the MNP control chart by Lu [40] contains a simulated data table of 30 samples in their article, where each sample is a single object that records the count of defects for three quality characteristics. Similarly, Chiu and Kuo [36] presented an exemplification of their MP control chart using a simulated data table of 26 samples, where each sample represents an object that records the D number of defects or non-conformities associated with three quality characteristics.

In the T2Qv control chart presented in this article, each of the individuals (rows) that make up the different samples can have different configurations depending on the categories of the variables.

On the other hand, other authors who have investigated multivariate control charts for attribute data, although they consider several quality characteristics in their analysis, ultimately classify each individual by only one of the analyzed variables. This is the case with Ranjan-Mukhopadhyay [41], whose proposal is demonstrated with an application case that controls 7 quality characteristics in 24 samples whose size varies between 20 and 404 individuals. The variables correspond to 6 types of paint defects on the cover of ceiling fans: poor coverage, overflow, puckering defect, bubbles, paint defects, and polishing defects. The seventh characteristic is the absence of defects. Each individual is classified by their most predominant defect; therefore, only one type of defect or absence of defects appears in their record, resulting in a loss of information about the combined effect of the variables on the process.

Control charts that incorporate multivariate techniques are applied to the analysis of categorical and numerical variables in the same data table. For example, the PCA Mix control chart [30] converts attribute variables into dummy variables and treats them together with continuous variables to generate a kernel matrix and calculate the principal components. One weakness of this proposal is that its performance decreases in the presence of extremely imbalanced proportions of qualitative variable

categories, a situation that is proposed to be corrected by applying the Kernel PCA (KPCA) method, a non-linear version of the conventional PCA that models non-Gaussian distributions.

These performance adjustments correspond to a phase II of control charts, with a view to their optimization. The T2Qv is a multivariate process control chart that handles qualitative variables in phase I; therefore, its efficiency evaluation has not yet been considered, so both charts are not comparable. However, it is precisely the change in the distribution of the variable categories, from balanced to imbalanced, or vice versa, that the T2Qv detects as an out-of-control point.

An interesting proposal is that of Saltos Segura *et al.* [45], who use the concept of depth for analyzing a data table in the field of education, but ultimately, the representation of academic performance is univariate and carried out using an r control chart. The ACM, being a factorial technique that works in terms of variability association, causes information common to all or the vast majority of cases to be stable and therefore not appear on the primary axes, at best it is concentrated at the graph's origin. In fact, in the T2Qv application, when a category of a variable is constant in a table of the database, an error is reported that prevents execution, but if there is at least one different case, it would be represented as a point far away at some extreme of the graph, while the category with the highest frequency would be located as a point at the center. This could be a characteristic of the nature of the ACM, which can only partially represent the variability of information in its two dimensions.

To correct this problem, the methodology proposed in this research, in conjunction with the ACM technique, uses the T2Qv chart which, as established in the sensitivity analysis, works best with the highest number of dimensions, i.e., collects the most variability to identify the out-of-control point (the point table). Additionally, a subsequent comparative analysis establishes the χ^2 distance between the values reported by the categories of the concatenated table and the point table for a specific variable and represents it in a bar chart, which allows the identification of the variables that are producing the most anomalies. Finally, the application facilitates the comparison of the distribution of categories of the variable analyzed in the point table and the concatenated table through an interactive graphical resource that uses pie charts.

An opportunity for future research related to multivariate control for qualitative variables could be the optimization of the graph, with a control limit that adjusts to the specific parameters of the analysis, taking it to a phase II of statistical process control. Another opportunity would be the development of a methodology that goes beyond the analysis of the first latent dimension, which is what the proposal of this research does when applying ACM. The incorporation of a Meta Biplot [20], for example, could be viable, a technique that cross-analyzes all latent dimensions.

7. Conclusions

This article presents a tool for multivariate statistical process control that performs analysis of qualitative data, which is called the T2Qv control chart, based on Multiple Correspondence Analysis. Normalized coordinates are represented by the robust Hotelling T2 chart.

To facilitate the dissemination and application of the proposed method, a reproducible statistical package has been developed in R, called T2Qv and available on CRAN, which allows for the visualization of results in a flat or interactive manner, and includes a Shiny panel that contains all integrated functions in the same space.

This proposal generates a graph of the MCA of the concatenated table, which serves as a reference for comparing other MCA graphs of tables that have been identified as out-of-control points on the Hotelling chart. This allows verification of which variable categories have had changes in their location in both graphs, which may be causing changes in the centrality of the process and causing it to be out of control.

To facilitate interpretation of the behavior of the variables, a chi-squared distance analysis is performed between the categories of the concatenated table and those of the tables reported as out of control, analytically and graphically, including interactive graphics that present the percentage distribution of the analyzed variable categories.

In a multivariate context, all variables contribute to a greater or lesser extent to explaining the behavior of the process, so that out-of-control output cannot be attributed to the individual action of a variable or to the separate action of a group of them, but rather to the combined effect of correlated variables. Therefore, a multivariate approach is necessary in statistical process control.

The sensitivity analysis determined that the T2Qv control chart has good performance when working with high dimensions, but loses stability at low dimensions.

There are not many published proposals for multivariate statistical process control for qualitative variables. The differences between procedures for determining statistics and control charts in this field make comparison difficult.

The T2Qv control chart addresses the need for a multivariate control chart for qualitative variables in social processes, where the use of nominal and ordinal variables is very common.

References

1. Gutiérrez, H.; de la Vara Salazar, R. *Control estadístico de la calidad y seis sigma*; Vol. 3, McGraw Hill Education, 2013; p. 152 – 253.
2. Ramos, M. Una alternativa a los métodos clásicos de control de procesos basada en coordenadas paralelas, métodos Biplot y Statis. PhD thesis, 2017.
3. Li, J.; Tsung, F.; Zou, C. Directional control schemes for multivariate categorical processes. *Journal of Quality Technology* **2012**, *44*, 136–154.
4. Hotelling, H. Multivariate quality control. Techniques of statistical analysis. *McGraw-Hill, New York* **1947**.
5. Lowry, C.A.; Woodall, W.H.; Champ, C.W.; Rigdon, S.E. A multivariate exponentially weighted moving average control chart. *Technometrics* **1992**, *34*, 46–53.
6. Roberts, S. Control chart tests based on geometric moving averages. *Technometrics* **2000**, *42*, 97–101.
7. Crosier, R.B. Multivariate Generalizations of Cumulative Sum Quality-Control Schemes. *Technometrics* **1988**, *30*, 291–303.
8. Page, E. Continuous inspection schemes. *Biometrika* **1954**, *41*, 100–115.
9. APARISI, F. Hotelling's T2 control chart with adaptive sample sizes. *International Journal of Production Research* **1996**, *34*, 2853–2862, [https://doi.org/10.1080/00207549608905062]. doi:10.1080/00207549608905062.
10. Aparisi, F.; Haro, C.L. Hotelling's T2 control chart with variable sampling intervals. *International Journal of Production Research* **2001**, *39*, 3127–3140, [https://doi.org/10.1080/00207540110054597]. doi:10.1080/00207540110054597.
11. Faraz, A.; Parsian, A. Hotelling's T2 control chart with double warning lines. *Statistical Papers* **2006**, *47*, 569–593. doi:10.1007/s00362-006-0307-x.
12. Ruiz-Barzola, O. Gráficos de Control de Calidad Multivariantes con Dimension Variable. PhD thesis, Universitat Politécnica de Valéncia, 2013.
13. Shabbak, A.; Midi, H. An improvement of the hotelling statistic in monitoring multivariate quality characteristics. *Mathematical Problems in Engineering* **2012**, *2012*.
14. Liu, Y.; Liu, Y.; Jung, U. Nonparametric multivariate control chart based on density-sensitive novelty weight for non-normal processes. *Quality Technology & Quantitative Management* **2020**, *17*, 203–215.
15. Xue, L.; Qiu, P. A nonparametric CUSUM chart for monitoring multivariate serially correlated processes. *Journal of Quality Technology* **2020**, pp. 1–14.
16. Mahalanobis, P. On the generalised distance in statistics. Proceedings of the national Institute of Science of India, 1936, Vol. 12, pp. 49–55.
17. Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. *Expert Systems with Applications* **2014**, *41*, 1701–1707. doi:10.1016/j.eswa.2013.08.068.
18. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **1901**, *2*, 559–572.
19. Gabriel, K.R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **1971**, *58*, 453–467.

20. Galindo-Villardón, P.; Vicente-Villardón, J.; Zarza, C.A.; Fernandez-Gómez, M.J.; Martın, J. JK-META-BIPLOT: una alternativa al método STATIS para el estudio espacio temporal de ecosistemas.
21. Benzecri., J. *OL'analyse des correspondances. En L'Analyse des Données: Leçons sur L'analyse Factorielle et la Reconnaissance des Formes et Travaux*; Paris - 1973, 1973.
22. des Plantes, L. Structuration des tableaux à trois indices de la statistique. *Université de Montpellier II, Thesis* **1976**.
23. Robert, P.; Escoufier, Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient. *Journal of the Royal Statistical Society Series C: Applied Statistics* **1976**, *25*, 257–265.
24. Lavit, C.; others. Présentation de la méthode STATIS permettant l'analyse conjointe de plusieurs tableaux de données quantitatives. *Les Cahiers de la Recherche Développement* **1988**, pp. 49–60.
25. Inselberg, B.; Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. Proceedings of the first IEEE conference on visualization: visualization90. IEEE, 1990, pp. 361–378.
26. Edwards, A.; Cavalli-Sforza, L. A method for cluster analysis. *Biometrics* **1965**, pp. 362–375.
27. Filho, D.; de Oliveira, L. Multivariate quality control of batch processes using STATIS. *The International Journal of Advanced Manufacturing Technology* **2016**, *82*, 867–875.
28. Ramos-Barberán, M.; Hinojosa-Ramos, M.V.; Ascencio-Moreno, J.; Vera, F.; Ruiz-Barzola, O.; Galindo-Villardón, M.P. Batch process control and monitoring: a Dual STATIS and Parallel Coordinates (DS-PC) approach. *Production & Manufacturing Research* **2018**, *6*, 470–493, [https://doi.org/10.1080/21693277.2018.1547228]. doi:10.1080/21693277.2018.1547228.
29. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Multivariate control chart based on PCA mix for variable and attribute quality characteristics. *Production & Manufacturing Research* **2018**, *6*, 364–384, [https://doi.org/10.1080/21693277.2018.1517055]. doi:10.1080/21693277.2018.1517055.
30. Ahsan, M.; Mashuri, M.; Wibawati.; Khusna, H.; Lee, M.H. Multivariate Control Chart Based on Kernel PCA for Monitoring Mixed Variable and Attribute Quality Characteristics. *Symmetry* **2020**, *12*. doi:10.3390/sym12111838.
31. Ahsan, M.; Mashuri, M.; Khusna, H. Comparing the performance of Kernel PCA Mix Chart with PCA Mix Chart for monitoring mixed quality characteristics. *Scientific Reports* **2022**, *12*, 1–12.
32. Ahsan, M.; Mashuri, M.; Kuswanto, H.; Prastyo, D.D.; Khusna, H. Outlier detection using PCA mix based T2 control chart for continuous and categorical data. *Communications in Statistics - Simulation and Computation* **2021**, *50*, 1496–1523, [https://doi.org/10.1080/03610918.2019.1586921]. doi:10.1080/03610918.2019.1586921.
33. Farokhnia, M.; Niaki, S.T.A. Principal component analysis-based control charts using support vector machines for multivariate non-normal distributions. *Communications in Statistics - Simulation and Computation* **2020**, *49*, 1815–1838, [https://doi.org/10.1080/03610918.2018.1506032]. doi:10.1080/03610918.2018.1506032.
34. Liu, Y.; Liu, Y.; Jung, U. Nonparametric multivariate control chart based on density-sensitive novelty weight for non-normal processes. *Quality Technology & Quantitative Management* **2020**, *17*, 203–215.
35. Holgate, P. Estimation for the bivariate Poisson distribution. *Biometrika* **1964**, *51*, 241–287.
36. Chiu, J.E.; Kuo, T.I. Attribute control chart for multivariate Poisson distribution. *Communications in Statistics-Theory and Methods* **2007**, *37*, 146–158.
37. Lee, L.H.; Costa, A.F.B. Control charts for individual observations of a bivariate Poisson process. *The International Journal of Advanced Manufacturing Technology* **2009**, *43*, 744–755.
38. Laungrungrong, B.; M, C.B.; Montgomery, D.C. EWMA control charts for multivariate Poisson-distributed data. *International Journal of Quality Engineering and Technology* **2011**, *2*, 185–211.
39. Epprecht, E.K.; Aparisi, F.; García-Bustos, S. Optimal linear combination of Poisson variables for multivariate statistical process control. *Computers & operations research* **2013**, *40*, 3021–3032.
40. Lu, X. Control chart for multivariate attribute processes. *International Journal of Production Research* **1998**, *36*, 3477–3489.
41. Ranjan-Mukhopadhyay, A. Multivariate attribute control chart using Mahalanobis D 2 statistic. *Journal of Applied Statistics* **2008**, *35*, 421–429.
42. Taleb, H.; Limam, M.; Hirota, K. Multivariate fuzzy multinomial control charts. *Quality Technology & Quantitative Management* **2006**, *3*, 437–453.

43. Taleb, H. Control charts applications for multivariate attribute processes. *Computers & Industrial Engineering* **2009**, *56*, 399–410.
44. Pastuizaca-Fernández, M.N.; Carrión-García, A.; Ruiz-Barzola, O. Multivariate multinomial T 2 control chart using fuzzy approach. *International Journal of Production Research* **2015**, *53*, 2225–2238.
45. Saltos Segura, G.; Flores Sánchez, M.; Horna Huaraca, L.; Morales Quinga, K. NEW METHODOLOGIES APPLIED TO MULTIVARIATE MONITORING OF STUDENT PERFORMANCE USING CONTROL CHARTS AND THRESHOLD SYSTEMS. *Perfiles* **2020**, *1*, 68–74.
46. López, C.P. *Técnicas de análisis multivariante de datos*; Pearson Educación, 2004.
47. Pearson, K. LIII.On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6* **1901**, *2*, 417 – 441. doi:10.1080/14786440109462720.
48. Hotelling, H. Analysis of a complex of statistical variables into principal components. **1933**. *24*, 417 – 441. doi:10.1037/h0071325.
49. Ch, S.; others. General intelligence objectively determined and measured. *American Journal of Psychology* **1904**, *15*, 201–293.
50. Thurstone, L.L. Multiple-factor analysis; a development and expansion of The Vectors of Mind. **1947**.
51. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
52. Michailidis, G.; Leeuw, J.D. The Gifi system of descriptive multivariate analysis. *Statistical Science* **1998**, pp. 307–336.
53. Escofier, B.; Pagès, J. Multiple factor analysis (AFMULT package). *Computational Statistics & Data Analysis* **1994**, *18*, 121–140. doi:https://doi.org/10.1016/0167-9473(94)90135-X.
54. Rojas-Preciado, W.; Rojas-Campuzano, M.; Galindo-Villardón, P.; Ruiz-Barzola, O. *T2Qv: Control Qualitative Variables*. R package version 0.1.0.
55. Ali, M.R.; Aslam, M. Design of control charts for multivariate Poisson distribution using generalized multiple dependent state sampling. *Quality Technology & Quantitative Management* **2019**, *16*, 629–650.
56. Jiang, W.; Au, S.; Tsui, K.L.; Xie, M. Process monitoring with univariate and multivariate c-charts. *Technical Report, the Logistics Institute, Georgia Tech, and the Logistics Institute-Asia Pacific* **2002**.